Correction rapide de l'examen de Modèle linéaire 2021/2022

Le barème est donné à titre indicatif.

Les trois exercices portent sur le même jeu de données concernant la région de Boston. Il contient les informations ci-dessous sur 506 quartiers de cette métropole.

Pour chacun des quartiers, on a :

- `crim` : taux de criminalité du quartier (nombre d'infractions criminelles pour 10000 habitants pour 1 an)
- `zn` : proportion de terrains résidentiels formés de propriétés de plus de 2300 mètres-carrés
- `indus` : proportion du quartier occupée par des zones industrielles
- `chas` : le quartier est-il en bordure de la rivière Charles (modalités : "`Yes`" ou "`No`")
- `nox` : concentration annuelle moyenne en oxyde d'azote (polluant, unité : pphm)
- `rm` : nombre moyen de pièces par logement
- `age` : proportion de logements occupés par leurs propriétaires et construits avant 1940
- `dis` : distance moyenne aux cinq principaux bassins d'emploi de Boston
- `rad` : temps d'accès moyen depuis un logement jusqu'aux grands axes de circulation (en minutes)
- `tax` : taux d'impôt foncier du quartier
- `ptratio` : rapport nombre d'élèves sur nombre d'enseignants dans les écoles du quartier
- `lstat` : proportion de la population de classe inférieure
- `logmv` : logarithme de la valeur moyenne (en milliers de dollars US) des logements

Exercice 1 (≈ 7 points)

Dans cet exercice, on souhaite prédire la variable $Y = \log(\texttt{crim})$, qui est le logarithme de la variable `crim`.

1.1 On introduit la covariable $X_1 = \texttt{rad}/10$ et on ajuste le modèle

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

Les résultats obtenus sont la table 1. Quelle est la part de variabilité de Y expliquée par X_1 ? Rappeler la définition de la quantité que vous utilisez pour répondre à cette question. Quelle quantité définie au niveau de la population estime-t-elle ?

> 73% de variabilité est expliquée...
> $R^2 = SSR/SST$, où $SST = \sum(y_i - \bar{y})^2$, $SSR = \sum(\hat{y}_i - \bar{y})^2$.
> Elle estime $R^2_{\text{pop}} = \beta_1^2 \text{Var}(X)/\text{Var}(Y) = \text{Cor}^2(X, Y)$.

1.2 Écrire R^2 en fonction de $\widehat{\sigma}^2$, l'estimation de la variance de l'erreur, et $\hat{s}_y^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \bar{y})^2$, l'estimation de la variance de Y.

> $R^2 = 1 - SSE/SST = 1 - (n-2)\widehat{\sigma}^2 \Big/ (n-1)s_y^2.$

1.3 Les distributions des variables Y et X_1 sont représentées dans la figure 1. Les résultats numériques de cette première régression sont dans la table 1. Comment s'interprète la valeur estimée de β_1 ? Les intervalles de confiances sont-ils utilisables avec ces informations ?

> β_1 : augmentation de la prédiction de Y (la log-criminalité) lorsque `rad` (temps d'accès grands axe) est augmenté de 10 unité.
> IC : la figure 1 ne permet pas de conclure

Table 1 – **Régression de la log-criminalité en fonction de** `rad`

	$\widehat{\beta}_i$	$\text{sd}(\widehat{\beta}_i)$	int. confiance (95%)
Intercept	-2.80	0.07	$[-2.95; -2.66]$
`rad`/10	2.12	0.06	$[2.00; 2.24]$

$R^2 \approx 0.73$ $\quad R^2_{\text{adj}} \approx 0.73$ $\quad \widehat{\sigma}^2 \approx 1.128$ $\quad \hat{s}^2_y \approx 4.674$

1.4 La figure 2 donne différentes représentations graphiques. Commentez rapidement ces graphiques, et complétez les réponses fournies précédemment.

> Figure 2 gauche : résidus presque gaussiens. Donc IC à peu près justes.
> Figure 2 milieu : peu de valeurs extrêmes, pas de courbe régulière
> Figure 2 droite : un gros trou dans les valeurs observées de `rad`/10...

1.5 Donnez une estimation du nouveau coefficient $\beta_{\texttt{rad}}$ si l'unique covariable du modèle est $\texttt{rad} = 10X_1$.

> $\beta_{\texttt{rad}} = \widehat{\beta}_1/10$ pour que le produit βx ne change pas.

1.6 On introduit la variable $X_2 = \texttt{lstat}/10$ et on s'intéresse au modèle

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 X_2 + \epsilon.$$

Mettez X_1 en facteur des deux termes où il apparait, et constatez que la quantité en facteur de X_1 (son effet) dépend linéairement de X_2.

> $$Y = \beta_0 + (\beta_1 + \beta_{1:2} X_2) X_1 + \beta_2 X_2 + \epsilon.$$

1.7 Les résultats numériques sont donnés dans la table 2. On admet que les hypothèses qui justifient les intervalles de confiance sont vérifiées. Commenter les valeurs estimées des coefficients β_1, β_2 et $\beta_{1:2}$. En particulier, comment comprenez-vous le signe de la valeur estimée de $\beta_{1:2}$?

Table 2 – **Régression multiple de la log-criminalité**

	$\widehat{\beta}_i$	$\text{sd}(\widehat{\beta}_i)$	int. confiance (95%)
Intercept	-3.93	0.14	$[-4.20; -3.66]$
`rad/10`	2.12	0.13	$[1.97; 2.50]$
`lstat/10`	1.14	0.11	$[0.92; 1.36]$
(`rad/10`) × (`lstat/10`)	-0.28	0.08	$[-0.44; -0.13]$
$R^2 \approx 0.79$		$R^2_{\text{adj}} \approx 0.79$	

$\beta_1 \approx 2 > 0$
$\beta_2 \approx 1 > 0$
$\beta_{1:2} \approx -0.3 < 0$

À X_2 fixé, l'effet de l'augmentation d'une unité de X_1 dépend de X_2 et vaut $\approx \beta_1 + \beta_{1:2} X_2$. Cet effet diminue à mesure que X_2 est grand ($\beta_{1:2} < 0$).

On peut aussi factoriser sous la forme

$$Y = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_{1:2} X_2) X_2 + \epsilon.$$

À X_2 fixé, l'effet de l'augmentation d'une unité de X_2 dépend de X_2 et vaut $\approx \beta_2 + \beta_{1:2} X_1$. Cet effet diminue à mesure que X_1 est grand ($\beta_{1:2} < 0$).

Exercice 2 (≈ 6 points)

On souhaite maintenant prédire $Y =$ `logmv` à l'aide des 12 autres variables du jeu de données. Les corrélations entre toutes les variables sont données dans la figure 3, à gauche.

2.1 La variable `chas` est catégorielle. Comment peut-on l'inclure dans un modèle linéaire ?

deux modalités $\implies$ une variable binaire (par exemple, =1 $\iff$ chas=Yes)

2.2 En utilisant la partie gauche de la figure 3, donnez la corrélation entre les variables `rad` et `tax`. Commentez.

0.91. Cette valeur est grande. Elle va perturber l'inférence par moindres carrés car présence de forte corrélation dans les covariables.

2.3 On note $\mathbb{X}$ la matrice du plan d'expérience (la première colonne est composée de $1, \dots$) et $\mathbb{Y}$ le vecteur colonne des valeurs observées de $Y =$ `logmv`. On note $\beta = (\beta_0, \beta_1, \dots, \beta_{12})$ le vecteur des effets et σ^2 la variance du terme d'erreur. En supposant que les erreurs sont gaussiennes, écrire la log-vraisemblance $\ell(\beta, \sigma^2)$ en fonction de ces quantités.

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left\|\mathbb{Y} - \mathbb{X}\beta\right\|^2 + \text{constante}$$

2.4 On pose $\beta_{-0} = (\beta_1, \ldots, \beta_{12})$ On s'intéresse à la méthode Lasso, où

$$\widehat{\beta}(\lambda) = \text{argmin}_\beta \left\|\mathbb{Y} - \mathbb{X}\beta\right\|_2^2 + \lambda\left\|\beta_{-0}\right\|_1.$$

À quoi sert-elle ?

A forcer l'annulation de coordonnées de l'estimation

2.5 Les différentes estimations $\widehat{\beta}(\lambda)$ et la valeur $\hat{\lambda}$, choisie par validation croisée, sont représentées à droite de la figure 3. À la l'aide de cette figure, répondre aux trois questions ci-dessous.

(a) Donnez le nom d'une variable v pour laquelle l'estimation $\widehat{\beta}_v(\widehat{\lambda})$ est nulle.

(b) Quelle est la variable v pour laquelle l'estimation $\widehat{\beta}_v(\widehat{\lambda})$ est la plus grande en valeur absolue ? Quel est son signe ?

(c) Quand λ augmente, le coefficient $\widehat{\beta}_{\texttt{tax}}(\lambda)$ en facteur de `tax` décroit vers 0 en valeur absolue, puis s'éloigne de 0 au moment où $\widehat{\beta}_{\texttt{rad}}(\lambda)$ s'annule. Que se passe-t-il ?

(a) rad, zn age
(b) lstat, signe <0
(c) ces deux variables sont fortement corrélées (voir quest 1). La disparition de rad est compensée. . .
+ zn qui s'annule et qui est négativement corrélé avec rad et tax (-0.31)

Exercice 3 (≈ 7 points)

Comme dans l'exercice 2, on souhaite prédire $Y =$ `logmv` à l'aide des 12 autres variables du jeu de données. **On suppose dans cet exercice que les 12 covariables numériques sont transformées pour être centrées-réduites.**

3.1 La méthode de sélection de variable rétrograde, partant du modèle complet, appliquée sur le critère BIC, donne les résultats de la table 3. La variable `nox` indique un niveau de pollution. Comment interprétez-vous les résultats quant à cette pollution ?

Table 3 – **Régression multiple de `logmv` obtenu par critère BIC**

	$\widehat{\beta}_i$	sd($\widehat{\beta}_i$)	int. confiance (95%)
Intercept	3.0345	0.0092	[3.0165; 3.0525]
`lstat`	−0.2300	0.0143	[−0.2581; −0.2019]
`ptratio`	−0.0768	0.0112	[−0.0988; −0.0548]
`dis`	−0.0854	0.0147	[−0.1142; −0.0566]
`nox`	−0.0795	0.0176	[−0.1141; −0.0449]
`rm`	0.0685	0.0120	[0.0449; 0.0921]
`chasYes`	0.0341	0.0094	[0.0157; 0.0526]
`tax`	−0.0403	0.0141	[−0.0679; −0.0127]

$R^2 \approx 0.75$ $\quad$ $R^2_{\text{adj}} \approx 0.74$

Toute chose étant égale par ailleurs, augmentation de la pollution fait baisser les prix (effet < 0, voir IC)

3.2 En partant du modèle complet à 12 covariables, la première variable éliminée dans cette méthode rétrograde est la variable `age`. Est-ce lié au choix du critère BIC ?

Non. Pour comparer des modèles de même dimension, c'est-à-dire à 11 covariables, on n'utilise que le R^2

3.3 Dans quel contexte utilise-t-on plutôt BIC que AIC ? Que peut-on dire de la covariable `zn` qui n'a pas été retenue par le modèle de la table 3, mais que le critère AIC propose de conserver ?

BIC = explicatif AIC = prédictif

`zn` : Ne gène pas, ou améliore un peu les qualités prédictives d'un modèle linéaire. Mais, sachant les autres covariables déjà incluses, l'influence de `zn` sur les prix n'est pas clairement établie...

3.4 La méthode progressive, appliquée au critère BIC, choisit le même modèle à 7 covariables que celui de la table 3. Cette égalité entre les modèles retenu par la méthode progressive et la méthode rétrograde est-elle attendue systématiquement ou est-ce un résultat spécifique à ce jeu de données ?

spécifique. On a vu des cas de différence en TP.

3.5 La première variable introduite par la méthode progressive est `lstat`. Peut-on le justifier avec la matrice de corrélation de la figure 3 ? (On attend ici une réponse argumentée.)

Dans la méthode progressive, pour la première covariable, on choisit la covariable avec le plus grand $R^2 = \text{Cor}^2(Y, X_i)$.

Toutes les autres cor avec logmv sont < 0.81 en valeur absolue.

3.6 On note $X_1, \ldots, X_7$ les variables aléatoires qui modélisent les 7 covariables de la table 3. Et Y celle qui modélise la réponse. Que vaut (justifier)

$$\mathbb{E}(Y|X_1 = x_1, \ldots, X_7 = x_7) \text{ ?}$$

Comment estimez-vous cette quantité pour des valeurs numériques de $x_1, \ldots, x_7$ données ?

$$\begin{aligned}
\mathbb{E}(Y|X_1 = x_1, \ldots, X_7 = x_7) &= \mathbb{E}(\sum_j \beta_j X_j + \epsilon | X_1 = x_1, \ldots, X_7 = x_7) \\
&= \sum_j \beta_j x_j + \mathbb{E}(\epsilon | X_1 = x_1, \ldots, X_7 = x_7) \quad \text{(linéarité)} \\
&= \sum_j \beta_j x_j + \mathbb{E}(\epsilon) \quad (\epsilon \text{ indep. } X_j) \\
&= \sum_j \beta_j x_j.
\end{aligned}$$

Estimation :

$$\sum_j \widehat{\beta}_j x_j.$$

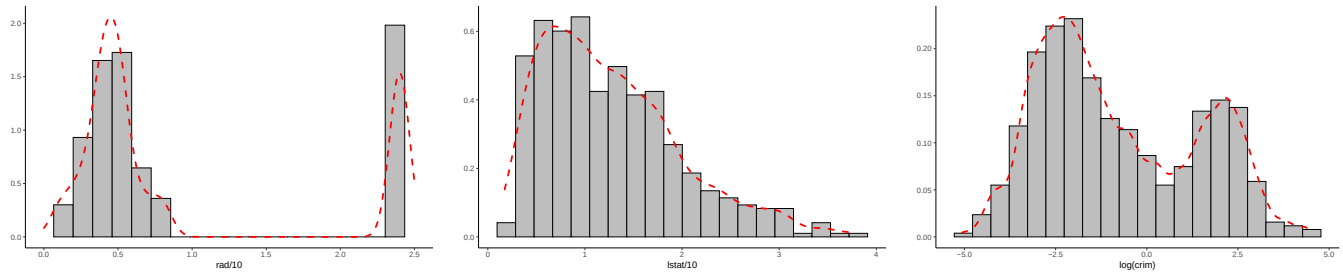

Figure 1 – **Distribution des variables** `rad/10` *à gauche*, `lstat/10` *au milieu* et $\log(\texttt{crim})$ *à droite*.

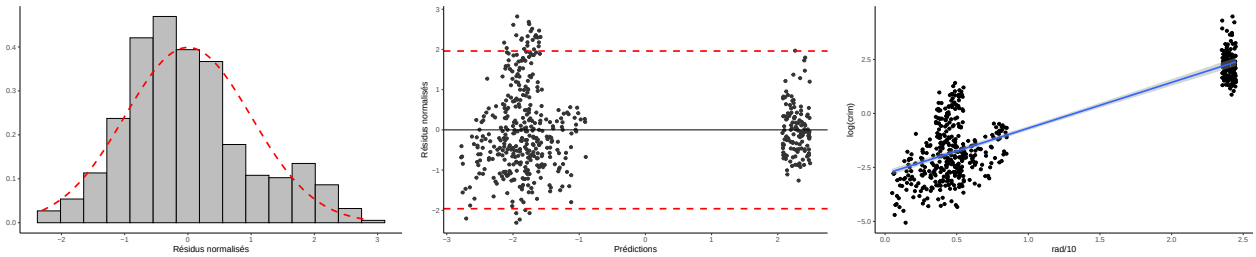

Figure 2 – **Étude du modèle linéaire simple** $\log(\texttt{crim}) = \beta_0 + \beta_1 \texttt{rad/10} + \epsilon$. *À gauche*, histogramme des résidus normalisés et densité de la loi normale centrée réduite en rouge. *Au milieu*, résidus normalisés en fonction de $\widehat{\beta}_0 + \widehat{\beta}_1 X_1$. *À droite*, $\log(\texttt{crim})$ en fonction de `rad/10`, et droite de régression ajustée sur les données.

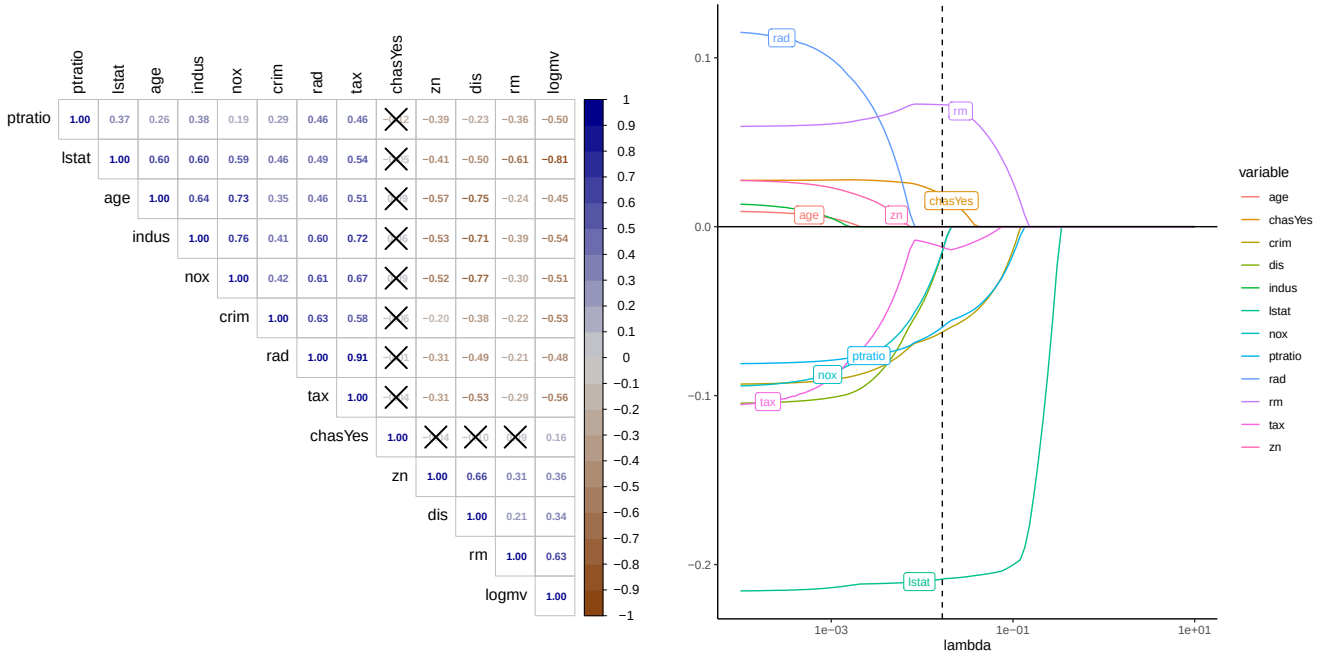

Figure 3 – *À gauche*, **matrice de corrélation** des variables du jeu de données. Les croix × représentent des corrélations non significativement différentes de 0.
À droite, représentation de l'estimation $\widehat{\beta}_i(\lambda)$ des coefficients par la **méthode Lasso** (sur l'axe des ordonnées) en fonction du paramètre de réglage λ (sur l'axe des abscisses). La droite verticale en pointillés correspond à la valeur $\widehat{\lambda}$ choisie par validation croisée.